11th DoD HPC Users Group Conference, June 2001, Biloxi, Miss.

Investigation, Visualization, and Interpretation of Large Scientific Data Sets

Dr. Brian Circelli and Mr. Paul Adams
U.S. Army Engineer Research and Development Center
Major Shared Resource Center
Vicksburg, Miss.

Dr. Joseph Werne, Dr. Michael Gourlay,
Dr. Christian Meyer, and Dr. Chris Bizon
Colorado Research Associates Division,
NorthWest Research Associates, Inc.
Boulder, Colo.

Abstract

The scientific investigation of physical phenomena in complex three-dimensional (3-D) flow fields is nontrivial and logistically challenging. Current supercomputer platforms, such as the Cray T3E and Origin 3000, are capable of simulating fluid flows with more than one billion mesh points, generating 32-bit floating-point files for individual flow fields several gigabytes in size. In order to develop an in-depth understanding of the flow morphology and underlying physical processes, researchers must overcome the difficulties created by the shear size of these massive data sets. Transferring data from one computer center to another, or even from one computer to another at the same site, presents a logistical challenge. In order to manage the interrogation of such large data sets, either automated postprocessing of the original floating-point files or size reduction for interactive investigation is necessary. Reduction of the data from floating-point to byte-scaled data is an option for scientific visualization, because the resulting files need only preserve sufficient resolution to satisfy viewing requirements, i.e., the reduced dynamic range must not be visually apparent. Furthermore, once the data have been scaled to a byte range from 0 to 255, compression algorithms (e.g., gzip) can accomplish additional reductions in file size. Visualization tools that take advantage of internal hardware texture mapping to render 3-D flow fields provide a means for rapid navigation and interpretation of the data. This paper examines some of the issues associated with the storage, transport, and investigation of large data sets; methods for effectively processing the data; and how the difficulties associated with investigating large data sets can be overcome.

Introduction

Within the past 5 years, the speed of supercomputers has grown immensely. The amount of data being generated by these computers has grown even more. At the U.S. Army Engineer Research and Development Center (ERDC) Major Shared Resource Center (MSRC), the amount of data being stored has grown to more than 80 terabytes. For example, the Airborne Laser II Challenge Project has generated more than 20 terabytes of the data archived at the ERDC MSRC. The analysis of data of this order of magnitude presents several difficulties that must be overcome.

Defining the Difficulties Associated with Large Data Sets

The first difficulty to overcome is the storage of the many terabytes of data resulting from time-accurate numerical simulations. The generated data sets can be on the order of hundreds of gigabytes for a small range of time-steps. Over an entire set of runs, this can cause the collection of data sets to grow to several terabytes. Despite the growth in the storage capacity of hard drives throughout the past 5 years, long-term storage of these files on hard drives that are local to the computer is impractical. The impracticality results from the storage capacity needs of other researchers who are using the same computational resources and are generating equally large data sets.

The methodology currently used to circumvent this problem is to transfer the large data sets over a high-speed network to another computer where the data will be archived and retrieved by the researcher at a later date. This mass storage facility computer normally contains enough disk space to store many gigabytes of data and files locally, until they can be migrated to robotic tape storage. Currently, the total disk storage capacity of the mass-storage-facility computer at the ERDC is 206 gigabytes.

The researcher’s second difficulty relates to the extraction of useful information contained within these massive data sets. Three-dimensional (3-D) scientific visualization is ideally used for initial qualitative understanding; but because of the inherent difficulties involved, it can be impractical for providing a researcher with his or her first impressions.

The visualization process usually begins with the retrieval of the archived data. For DoD Challenge projects spanning geographically

separated supercomputer centers, transferring massive data sets between sites can be a challenge. Indeed, even the transfer between computers within the same institution presents difficulties. For example, a Gigabit Ethernet network theoretically can transfer one billion bits per second. In reality, the number of bits transferred is closer to 800 million per second, or 100 megabytes per second. At 100 megabytes per second, it would take approximately 2.8 hours to transfer a one terabyte file. Unfortunately, Gigabit Ethernet networks currently exist only locally within individual centers. At the ERDC MSRC, the connection to the outside world occurs over a fiber optical connect, OC-12 line with the ability to transfer data at a rate of 622 megabits per second. This increases the transfer time of a terabyte file to about 3.6 hours. Furthermore, the Ethernet network at a researcher's home institution likely introduces more severe restrictions. A local site's Ethernet network capable of only transferring data at a rate of 100 megabits per second would significantly throttle down the rate of data transfer. A transfer rate of 100 megabits per second would require approximately 22.2 hours transferring a one terabyte file. This example neglects the impact of other researchers using the Ethernet network or the network interface cards.

A researcher may think waiting nearly a day to transfer a one terabyte file is reasonable; however, it is important to realize that even using striping, the hard drives of a computer have a finite rate of data transfer and can only transfer so much data at one time. Of even greater concern to a researcher is the realization that the computer used for file storage has a finite storage capability and may not have the necessary disk space storage available. Increasing disk space storage is a costly endeavor. For example, a terabyte of disk space currently costs approximately $100,000.

Proposed Alternatives to Manipulating Large Data Sets

The difficulties described above leave the researcher with two choices. The first choice is for the researcher to retrieve data from archival storage to the supercomputer where data post-processing analysis and visualization are performed using a Client/Server visualization package. One such available software visualization package is Ensight Gold, which can run on the researcher's local machine and connects to the Ensight Gold Server on the supercomputer. The main advantage of this choice is that the data remain local to the supercomputer, which provides disk space on the order of a terabyte. Unfortunately, that disk space

is usually of a temporary nature, and files older than a certain timestamp are likely to be purged if the temporary file system becomes saturated. An additional problem associated with this choice is that since the disk space on the supercomputer is shared among all users, the requisite space needed to postprocess and visualize the data may not be available, which again could trigger the automatic purging of files if the file system is saturated.

The second choice is to retrieve data from archival storage, transfer it to the supercomputer where postprocessing data analysis is performed, byte-scale the postprocessed data, and then transfer the byte-scaled data back to the researcher's home-site computer for visualization. This option provides the researcher with a more tangible method of manipulating such massive data sets. The term byte scaling is used to describe the process of converting floating-point data to an integer-data value between 0 and 255, where each data point is one byte of data. For example, suppose the results of postprocessing the raw data archived on mass storage and transferred to the supercomputer consisted of 32-bit floating-point numbers. The successful completion of byte scaling the postprocessed data would reduce the size of a 32-bit floating-point number to an eight-bit integer, which is exactly one byte of data. As illustrated in this example, the obvious benefit of byte scaling data is that it can reduce the data set size by a factor of four. Unfortunately, the process of byte scaling the original floating-point numbers also results in a compression of the dynamic range to 1/256, so this procedure is really only feasible for visualization purposes. This loss in qualitative resolution is an obvious disadvantage of the byte-scaling process; however, if the scaling is performed judiciously, the potential masking of valuable information can be mitigated.

Narrowing the visualization to a selected region of interest within the flow field may further reduce the data-file size. Here, the intent is to make the data set even more manageable by further reducing the data set size, without causing any further degradation in the quality of the data. The ERDC MSRC is currently employing this technique in support of the Airborne Laser II Challenge Project, which involves the 3-D simulation of turbulence in the lower stratosphere and upper troposphere. Postprocessed floating-point data files of size 2.8 gigabytes each have been reduced by a factor of four, to 700 megabytes by byte-scaling and then by an additional factor of three, to 233 megabytes by concentrating on the middle third of the computational domain. For these simulations, all of the fluid turbulence is contained within the middle third of the data volume. From this factor of 12 reduction in file size, the data files are then compressed using

gzip, producing final file sizes that are typically 70 to 80 times smaller than the original full floating-point-data volumes.

Scientific Visualization of Large Data Sets

The efficient visualization of large data sets requires software tools that take advantage of the advanced hardware architecture design built into computers intended for viewing large-scale scientific data. Specifically, 2- and 3-D internal hardware texture mapping provides a rapid capability of manipulating and moving through large rendered data-set volumes. This capability is vital to the researcher whose focus is on the timely interpretation of data and communication of the results to the scientific community.

One such tool that satisfies this need and is currently being used as a large-scale data volume renderer at the ERDC MSRC is a software visualization package named Ogle. Ogle is a 3-D vector and scalar scientific visualization tool based on OpenGL. Although Ogle is currently a developing research tool, it is well documented and supported, easy to use, and provides multifunctional capabilities including rendering data-set volumes, locating streamline paths, plotting vector field arrows, and plotting isosurfaces. Ogle also possesses the capability of reading compressed (zipped) input data files, i.e., Ogle automatically detects and decompresses data files on the fly. Aside from saving disk space, using a compressed data file is faster than reading the uncompressed data file, since the speed associated with uncompressing a file is usually faster than the speed associated with disk IO. This is especially true for disk IO across a network. During the course of the postprocessing and data analysis work performed at the ERDC MSRC, an additional factor of five to six in disk storage space savings was achieved by working only with compressed data sets, i.e., the size of compressed, byte-scaled data for a narrowed region of interest was reduced to approximately 55 megabytes, or just 2 percent of the size of the original data set of 2.8 gigabytes.

The multifunctionality of Ogle was one of the main reasons why the ERDC MSRC selected it as its current large-scale data volume-rendering tool. Unfortunately, Ogle uses only 2-D internal hardware texture mapping and does not take full advantage of current 3-D internal hardware texture mapping technology. For very large data volumes, this deficiency significantly increases the time necessary to perform volume renderings. However, plans are

currently under way to incorporate 3-D internal hardware texture mapping capabilities, as well as other capability upgrades.

ERDC MSRC Visualizations

During the past 6 months, the ERDC MSRC has been working in close collaboration with the Airborne Laser II (ABL) and Wake Turbulence (WT) Challenge Projects. The ABL/ERDC MSRC collaborative effort has involved postprocessing analysis and byte scaling of more than six terabytes of raw data that were generated by a 3-D pseudo-spectral flow solver. In support of the ABL effort, the ERDC MSRC Scientific Visualization Center (SVC) staff has consumed more than 50,000 CPU hours on the Naval Oceanographic Office (NAVO) MSRC and ERDC MSRC Cray-T3E supercomputers. Some of the results of this work were presented earlier during the Computational Fluid Dynamics (CFD) Session C of this conference in the form of 3-D scientific animations of the breakdown of a Kelvin-Helmholtz (KH) vortex. Figures 1(a-c) illustrate vortex tube behavior that is representative during the course of the KH vortex breakdown. Of particular interest is the disappearance, with time, of the coherent elongated vortex tubes that exist at earlier times in the flow. The identification and visualization of these structures would be difficult without the volume-rendering capabilities of Ogle.

The WT/ERDC MSRC collaborative effort has primarily involved the animation and scientific visualization of turbulent late wakes in density stratified flows. Unlike the ABL Challenge Project, the ERDC MSRC SVC staff was not asked to perform any postprocessing analysis or byte scaling of the raw data generated by the WT spectral flow solver. Instead, focus was directed at providing 3-D scientific animations to assist the WT researcher in the understanding and interpretation of the data generated from the large-scale direct numerical simulations. Some of the scientific animations generated by the ERDC MSRC SVC staff in support of the WT Challenge Project will be presented during the Challenge Projects Session C of this conference. Figures 2(a-c) illustrate the 3-D representation of the coherent pancake vortices that exist in a zero momentum density stratified flow. Of particular interest is the 3-D behavior of the pancake vortices as a function of time, as indicated by the streamtubes that create the ring-like coherent structures within the flow. Again, the identification and visualization of these structures would be difficult without the volume-rendering and streamtube capabilities of Ogle.

Conclusion

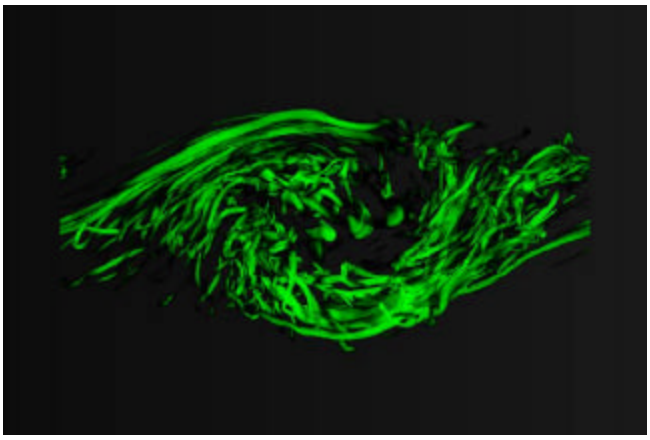
This paper has addressed some basic issues associated with very large scientific data sets. It has identified some fundamental difficulties typically encountered and has provided some methods for solving these problems. Specifically, the difficulties include the storage, transport, analysis, visualization, and interpretation of the information contained within the data sets. A solution to these difficulties is to find ways of reducing the large-scale data sets to a manageable size without severely degrading the dynamic range. Techniques such as byte scaling, focusing only on a subset of the overall flow domain, and data-set compression were presented as reasonable methodologies for creating more manageable data sets. For the analysis presented in this paper, a data-set size for a given run was reduced from approximately 2.8 gigabytes to 55 megabytes, which represents a savings in disk space of nearly 98 percent.

This paper has also addressed the visualization of very large data sets and the desirability of using visualization software that takes advantage of internal hardware texture mapping to help the researcher understand and interpret the information contained in the data sets. It has introduced a scientific visualization tool called Ogle that satisfies some of the researcher's current needs.

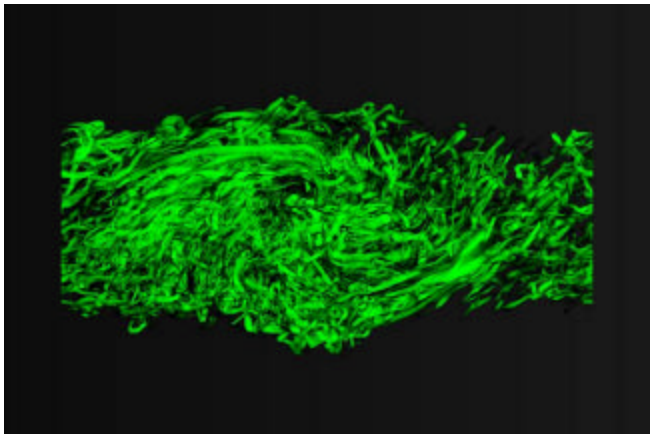
Finally, this paper, in concert with two other papers that have been presented at this conference, demonstrates what can be accomplished when the Challenge researchers and an MSRC join together to leverage off the strengths of each organization.

Acknowledgment

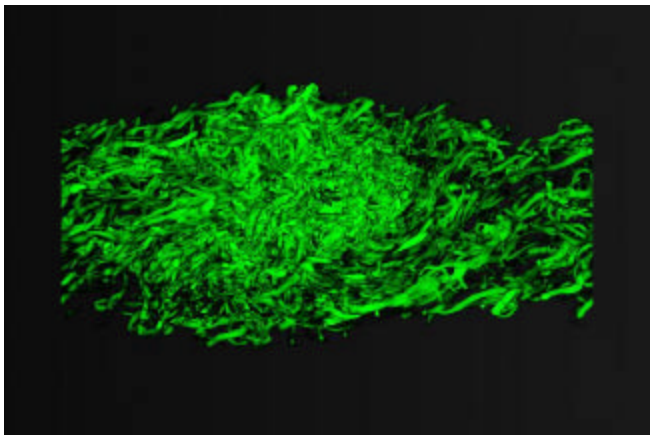
This work was supported in part by a grant of computer time from the DoD High Performance Computing Modernization Program at the ERDC MSRC, Vicksburg, Miss.



(a)

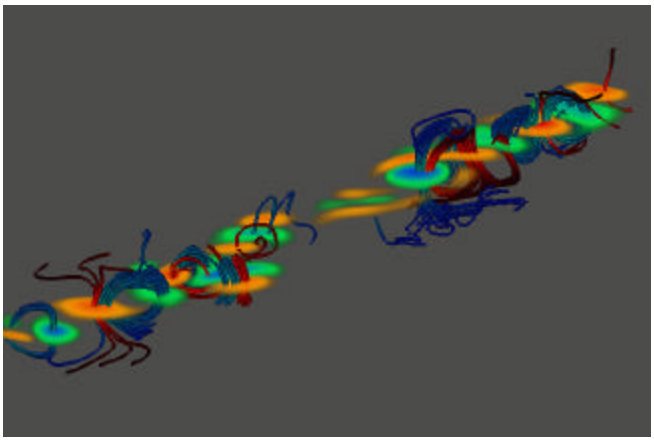


(b)

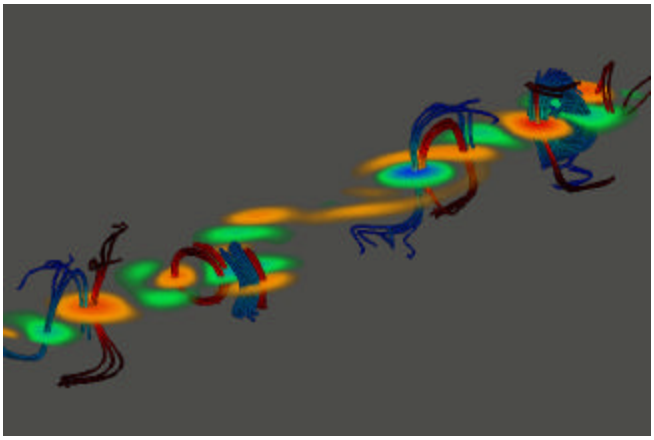


(c)

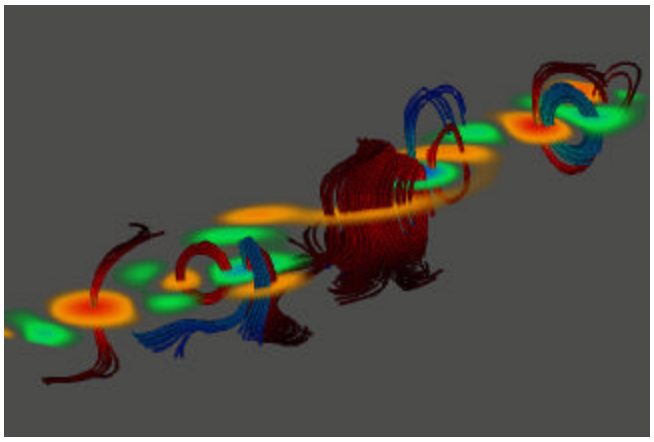
Figures 1(a-c). 3-D volume-rendered visualizations of the vortex tube field at three instances in time during the breakdown of a KH vortex billow in stratified turbulent shear flow.



(a)



(b)



(c)

Figures 2(a-c). 3-D volume-rendered visualizations of coherent pancake vortices in a zero momentum density stratified flow, at three instances in time. The generated streamtubes indicate the existence of coherent ring-like structures within the flow field.